Homework III

Due date: Sunday, Mar 17th 2024 at 11:59pm

Note: For written questions, you can turn in either a scanned copy of your handwritten answers or a PDF file of your answers. For programming questions, you need to submit your code and **also answer the asked questions in pdf**. Name the submission package as hw3_Lastname_Firstname.zip.

Problem 1: kNN Classifier (30 points)

You are asked to build a k-Nearest Neighbor (kNN) classifier. Here we use the heart data set. More information about the data can be found here[1]. The data set is included in hw3_datasets.zip on Canvas. In the heart data folder, there are three files: "trainSet.txt", "trainLabels.txt", and "test.txt". Each row of "trainSet.txt" corresponds to a data point whose class label is provided in the same row of "trainLabels.txt". Each row of "testSet.txt" corresponds to a data point whose class label needs to be predicted. Code to load data is given in "p1.py". You will train a classification model using "trainSet.txt" and "trainLabels.txt", and use it to predict the class labels for the data points in "testSet.txt".

1) Use the leave one out cross validation on the training data to select the best k among $\{1, 2, \ldots, 10\}$. Report the averaged leave-one-out error (averaged over all training data points) for each $k \in \{1, 2, \ldots, 10\}$.

2) Based on 1), use the best k to predict the class labels for test instances. You should also report the predicted labels for the testSet.

Problem 2: PCA (50 points)

You are asked to build a k-Nearest Neighbor (kNN) classifier based on dimenationality reduced data by PCA. The data set for evaluation is the gisette data set. More information about the data can be found here[2]. The data set is included in hw3_datasets.zip on Canvas. The data is in the same format as that in Problem 1. You will train a classification model using "trainSet.txt" and "trainLabels.txt", and use it to predict the class labels for the data points in "testSet.txt". A demo code is given in "p2.py", which includes data loading, model training, and visualization.

1) Train a kNN based on the original features. You should conduct cross-validation (of your choice) to select the best k. Describe the cross-validation approach, e.g., how many folds, how did you split the data. If you use a library for cross-validation, describe how the library does it. Report the best value of k under your cross-validation approach, and then report the testing accuracy corresponding to the best k.

[1] `https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)`

[2] `https://archive.ics.uci.edu/ml/datasets/Gisette`

2) Conduct PCA on the data and then learn a kNN model using the dimensionality reduced data. You should conduct cross-validation (of your choice) to select the best k and best d, where d is number of dimensions after PCA. Describe the cross-validation approach and report the best value of k and best value of d. Report the testing accuracy corresponding to the best k and best d.

3) Visualize the data distribution in the PCA projected space with d=1, d=2, and d=3, respectively. Include the visualization as figures in your pdf. Are there any visible clusters or patterns in the PCA plot? How does the PCA visualization change with different numbers of components?

Problem 3: (20 points)

Given a set of observations $\{x_1, x_2, ..., x_n\}$, where $x_i \in \{1, 2, ..., K\}$. Assume that each $x_i, i = 1, ..., n$ follows an identical and independent distribution specified by $Pr(x_i = k) = p_k$, where $\sum_{k=1}^{K} p_k = 1$. Please derive the maximum likelihood estimation of the parameter $p = (p_1, p_2, ..., p_K)$. Hint: You can use a one-vs-rest strategy. For example, when you consider p_1, x_i can be viewed as either 1 or not 1.